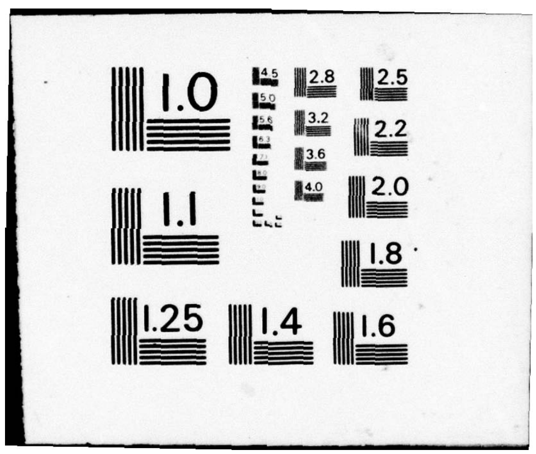
AD-A033 385 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF MATHEMATICS F/G 12/1
INFINITELY DIVISIBLE DISTRIBUTIONS IN STATISTICAL INFERENCE: HE--ETC(U)
NOV 76 S L SCLOVE AF-AFOSR-3050-76

UNCLASSIFIED AFOSR-TR-76-1258 NL

1 OF 1
ADA033385

END
DATE
FILMED
2 - 77


1.0
4.5
5.0
5.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

D D C
RECEIVED
DEC 10 [illegible]
A

Approved for public release;
distribution unlimited

(4)

INFINITELY DIVISIBLE DISTRIBUTIONS IN STATISTICAL INFERENCE: HEAVY-TAILED DISTRIBUTIONS AND CONVOLUTION MODELS

by

Stanley L. Sclove

Department of Mathematics
University of Illinois at Chicago Circle

November 1, 1976

PREPARED UNDER GRANT AFOSR 76-3050
FOR
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release;
distribution unlimited.

DDC
RECEIVED
DEC 10 1976
A

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
CHICAGO, ILLINOIS 60680

see 1473

INFINITELY DIVISIBLE DISTRIBUTIONS IN STATISTICAL INFERENCE:
HEAVY-TAILED DISTRIBUTIONS AND CONVOLUTION MODELS

Stanley L. Sclove

Department of Mathematics
University of Illinois at Chicago Circle

Outline

1. Introduction
2. Heavy-tailed distributions
3. The convolution model
4. Testing for normality
5. Testing for normal-plus-Poisson
6. Testing for normality, given normal-plus-Poisson
7. Testing for a Poisson distribution

BY
DISTRIBUTION/AVAILABILITY CODES
A

INFINITELY DIVISIBLE DISTRIBUTIONS IN STATISTICAL INFERENCE: HEAVY-TAILED DISTRIBUTIONS AND CONVOLUTION MODELS [1]

BY STANLEY L. SCLOVE

University of Illinois at Chicago Circle

Abstract

The family of infinitely divisible distributions is shown to provide alternative formulations in several inferential situations. In particular, the family provides heavy-tailed distributions and distributions for use in models involving convolutions, such as signal-plus-noise models. Characterizations of sub-families of the infinitely divisible family are used to obtain statistical tests of membership in those sub-families. Special attention is given to the normal and normal-plus-Poisson sub-families.

1. Introduction. A random variable (r.v.) is said to be *infinitely divisible (inf. div.)* if for every $n = 2,3,\ldots$ there exist independent identically distributed r.v.'s $X_{n1}, X_{n2}, \ldots$ such that the distribution of $X_{n1} + X_{n2} + \ldots + X_{nn}$ is the same as that of X. Thus, letting $U \overset{d}{=} V$ mean that U and V have the same distribution, such an X can be expressed in terms of a triangular array.

$$\begin{aligned} X &\overset{d}{=} X_{21} + X_{22} \\ &\overset{d}{=} X_{31} + X_{32} + X_{33} \\ &\;\;\vdots \\ &\overset{d}{=} X_{n1} + X_{n2} + \ldots + X_{nn} \\ &\;\;\vdots \end{aligned}$$

In terms of the characteristic function (c.f.) $f(u)$ of the r.v. X, this is equivalent to saying that for each $n = 2,3,\ldots$ there exists a c.f. $f_n(u)$ such that

$$f(u) = [f_n(u)]^n.$$

[1] Research supported by AFOSR Grant #76-3050.

AMS 1970 subject classifications. Primary 62E10; Secondary 62A99

Key words and phrases. Infinite divisibility, normality, Poisson distribution, cumulants, convolutions.

It will be shown that the family of inf. div. distributions is a family which can be useful in a number of models, especially those requiring heavy-tailed distributions and those in which the observed variable is a convolution, a sum of two independent r.v.'s.

The family of inf. div. distributions is quite broad. It includes the normal distributions, as well as the gamma distributions, the related exponential and chi-square distributions, and the double-exponential distributions. It includes the Poisson and compound Poisson distributions (distributions generated by putting a distribution on the Poisson parameter), though here we shall be interested primarily in continuous inf. div. distributions. The family also includes the generalized Poisson distributions (distributions of r.v.'s which are sums of a Poisson-distributed number of identically distributed r.v.'s). Some distributions which are *not* inf. div. are those with bounded support, those whose c.f. vanishes at some point on the real line, and those whose c.f. is an entire function which vanishes at some point in the complex plane.

Another way to demonstrate the breadth of the inf. div. family is to note that each of the following families of distributions contains the preceding: normal distributions, stable distributions, self-decomposable distributions, inf. div. distributions.

Relatively recently a number of researchers have used *stable* distributions for modelling various phenomena. Since every stable distribution is inf. div., the inf. div. distributions can be used wherever stable distributions are, and the result is a model which is less restrictive and can be valid under more general circumstances.

2. Heavy-tailed distributions. A number of researchers have studied stable distributions [see. e.g., DuMouchel (1973, 1975), Fama and Roll (1971)] because they are "heavy-tailed." A primary motivation for such studies is the observation of economists that the distributions of changes in stock prices seem to be rather heavy-tailed.

Stable distributions are indeed heavy-tailed. In fact, the only stable distribution with finite variance is the normal distribution. It is acknowledged [see, e.g., DuMouchel (1973), p. 469] that it is not necessary to use infinite-variance distributions in order to provide heavy-tailed distributions.

If fact, all infinitely divisible distributions are heavy-tailed. For, as will be shown below, their fourth cumulant, κ_4, is necessarily non-negative. Thus, letting μ_r denote the r-th central moment, we have

$$\begin{aligned} 0 &\leq \kappa_4 \\ &= \mu_4 - 3\mu_2^2 \\ &= \int(x-\mu)^4 dF(x) - 3\sigma^2 \int(x-\mu)^2 dF(x) \\ &= \int [(x-\mu)^4 - 3\sigma^2(x-\mu)^2] dF(x), \end{aligned}$$

where $F(x)$ denotes the distribution function and $\sigma^2 = \mu_2$ is the variance. Thus $dF(x)$ weights relatively heavily those points x for which

$$(x-\mu)^4 - 3\sigma^2(x-\mu)^2 \geq 0;$$

i.e.,

$$|x-\mu| \geq \sqrt{3}\sigma .$$

Thus $F(x)$ corresponds to a relatively heavy-tailed distribution. More precisely, a normal distribution has $\kappa_4 = 0$, so a distribution with positive fourth cumulant is heavy-tailed relative to the normal.

It remains to show that the fourth cumulant is non-negative.

THEOREM 2.1. *If* X *is inf. div. with finite fourth moment, then the fourth cumulant* κ_4 *is non-negative.*

PROOF. [Cf. Pierre(1969), p. 320.] Since X has finite fourth moment, the fourth cumulant exists and is given by

$$\kappa_4 = (d^4/du^4)\log f(u)|_{u=0}, \tag{2.1}$$

where $f(u)$ is the c.f. of X. Since X is inf. div., the logarithm of $f(u)$ can be written [see. e.g., Loève(1963), p. 293] as

$$\text{lof } f(u) = iu\mu + \int[\exp(iux) - 1 - iux]x^{-2}dK(x), \tag{2.2}$$

where $K(x)$ is monotone increasing and of bounded variation, $K(-\infty) = 0$, $K(\infty) = \text{Var}(X) < \infty$, and $\mu = E(X)$. The integrand is defined at the origin by continuity.

Now, from Loève(1963), p. 293, one sees that

$$-(d^2/du^2)\log f(u) = \int \exp(iux)dF(x) . \tag{2.3}$$

Thus the left-hand side of (2.3) is a c.f. Since this c.f. is twice differentiable, its second derivative is given by [Loève(1963), p. 200]

$$(d^4/du^4)\log f(u) = \int x^2\exp(iux)dK(x) .$$

Thus, by (2.1),

$$\kappa_4 = \int x^2 dK(x) \geq 0 . \tag{2.4}$$

3. The convolution model. Now consider the model

$$X = Y + Z, \tag{3.1}$$

where the r.v.'s Y and Z are independent, non-identically distributed and individually not observed. The r.v. Y has a distribution in a parametric family $\{P_\theta\}$ and Z has a distribution in another parametric family $\{Q_\omega\}$. Thus the family of distributions for the observed r.v. X is a family of convolution distributions $\{F_{\theta,\omega} = P_\theta * Q_\omega\}$, where $*$ is the convolution operation of distribution functions. This model is considered by Sclove and Van Ryzin (1971). They show that, when Z is discrete, maximum-likelihood estimation of the parameters θ and ω becomes intractable and that quite generally the method of moments offers a solution.

Any signal-plus-noise model is of the form (3.1). The model with discrete signal Z can occur in any counting process where the count is recorded as a measured electrical pulse which results from the actual count plus an error introduced by electrical noise in the counting mechanism. Another application arises in the problem of estimating the mean density of viruses (or bacteria) in a homogeneous solution where the "count" X is measured as the *area* on a slide occupied by the viruses where each virus occupies a unit of area. Hence, the total area is $Y + Z$ where Z is the number of viruses and Y is the sum of the deviations from the ideal (one virus per unit area) plus error in the measurement of area. In many such applications it is reasonable to take the distribution of the continuous variable Y to be normal and that of the discrete variable Z to be Poisson or Poisson-related (e.g., negative binomial or some other compound Poisson distribution).

Now, if $X = Y + Z$, where Y is normal and Z is Poisson or compound Poisson, then X is inf. div. In fact, the r.v. X is inf. div. if and only if (2.2) holds. The integrand in (2.2) is defined by continuity at the origin. Hence, since the limit as x tends to zero of the integrand $[\exp(iux) - 1 - iux]x^{-2}$ is $-u^2$, we have

$$\log f(u) = iu\mu - u^2\delta^2 + \int[\exp(iux - 1 - iux]x^{-2}dM(x) , \tag{3.2}$$

where δ^2 is the jump of $K(x)$ at the origin and $M(x)$ has no mass at the origin. This is equivalent to $X \overset{d}{=} \mu + Y + Z$, where Y has log c.f. equal to $-u^2\delta^2$ and hence is normal [with $\mathrm{Var}(Y) = 2\delta^2$], and Z has lof c.f. equal to the integral in (3.2). The r.v. Z is called the "Poisson component" of X or is said to be of "Poisson type." Thus the convolution model (3.1) with Z suitably distributed leads to an inf. div. X. Conversely, every inf. div. X obeys a convolution model.

It was noticed by Borges (1966) and later independently by Pierre (1967) that nullity of the fourth cumulant characterizes the normal distribution in the class of inf. div. laws.

THEOREM 4.1. *An inf. div. distribution is normal if and only if its fourth cumulant is zero.*

PROOF. If X is normal, then its c.f. $f(u)$ satisfies

$$\log f(u) = iu\mu - u^2\delta^2,$$

so that in the expansion of $\log f(u)$ all terms of degree greater than two vanish. The r-th cumulant is the coefficient of $i^r u^r/r!$ in the expansion. Hence all cumulants of order greater than two vanish. Conversely, suppose X is inf. div. and has zero fourth cumulant. By (2.2) and (3.2), X is normal if and only if $K(x)$ increases only at $x = 0$. Hence it suffices to prove that $K(x)$ increases only at $x = 0$. But $\kappa_4 = 0$, so this is immediate from (2.4).

Using this characterization of the normal distribution among inf. div. laws, one can construct a test of the hypothesis that an inf. div. r.v. is normally distributed. The hypothesis to be tested is

H: X is normally distributed, given that X is inf. div.

The alternative hypothesis is not H: X is inf. div. but is not normally distributed. The hypothesis is equivalent to $\kappa_4 = 0$.

An(unbiased) estimator for κ_4 is [Kendall and Stuart, p. 281, (12.29)]

$$k_4 = n^2[(n+1)m_4 - 3(n-1)m_2^2]/[(n-1)(n-2)(n-3)],$$

where

$$m_r = \Sigma_{i=1}^n (x_i - \bar{x})^r/n$$

is the sample analogue of μ_r (and is a biased estimator of μ_r). The statistic $k_4/\sigma(k_4)$, where $\sigma(S)$ denotes the standard deviation of the statistic S, is asymptotically normally distributed under H; so is $k_4/s(k_4)$, if $s(k_4)$ is a consistent estimator for $\sigma(k_4)$. The variance of k_4 is relatively complicated [Kendall and Stuart (1969), p. 290, (12.37)], but under the hypothesis of normality it reduces simply to

$$\sigma^2(k_4) = 48\sigma^8/(n-1)^3$$

[Kendall and Stuart (1969), p. 296, (12.71)]; a consistent estimator for this is obtained by substituting $s^2 = \Sigma_{i=1}^{n}(x_i-\bar{x})^2$ for σ^2. (Alternatively, one could replace s^8 by an unbiased estimator for σ^8.) At level α one rejects the hypothesis of normality if

$$|k_4/s(k_4)| > z(\alpha/2),$$

where $z(p)$ denotes the upper p-th percentage point of the standard normal distribution.

5. Testing for normal-plus-Poisson. Of special interest in the convolution model is the case in which Y is normal and Z is Poisson. Sclove and Van Ryzin (1971) give parameter estimates for such special cases. Before applying their results, it would be desirable to test the adequacy of such a special model. The adequacy of the normal-plus-Poisson assumption can be tested against the underlying assumption that X is inf. div. For this we need the following theorem [Pierre (1971), p. 348].

THEOREM 5.1. *Suppose* X *is inf.div. Then* $X \overset{d}{=} Y + Z$, *where* Y *is normal and* Z *is ordinary Poisson, if and only if*

(5.1) $$\kappa_6 - 2\kappa_5 + \kappa_4 = 0.$$

PROOF. The c.f. of a Poisson r.v. with parameter μ has logarithm equal to

(5.2) $$\mu[\exp(iu) - 1] .$$

The terms

$$iu\mu + \int[\exp(iux) - 1 - iux]x^{-2}dM(x)$$

of (3.2) are of the form (5.2) if and only if $M(x)$ increases only at $x = 1$. Now suppose (5.1) holds. Then

$$\begin{aligned} 0 &= \kappa_6 - 2\kappa_5 + \kappa_4 \\ &= (d^6/du^6)\log f(u)|_{u=0} - 2(d^5/du^5)\log f(u)|_{u=0} + (d^4/du^4)\log f(u)|_{u=0} \\ &= \int x^4 dM(x) - 2\int x^3 dM(x) + \int x^2 dM(x) \\ &= \int x^2(x^2 - 2x + 1)dM(x) \\ &= \int x^2(x-1)^2 dM(x) . \end{aligned}$$

Thus $M(x)$ increases only at $x = 1$ [an increase at $x = 0$ having already been excluded in replacing $K(x)$ by $M(x)$], and so Z is ordinary Poisson. Conversely, if $X \overset{d}{=} Y + Z$, where Y is normal and Z is Poisson with parameter μ, then the log c.f. is given by

$$\begin{aligned} \log f(u) &= -u^2\delta^2 + \mu[\exp(iu) - 1] \\ &= iu\mu + i^2u^2(2\delta^2+\mu)/2 + i^3u^3\mu/3! + i^4u^4\mu/4! + \dots \end{aligned}$$

so that

(5.3) $$\kappa_1 = \mu, \quad \kappa_2 = 2\delta^2 + \mu, \quad \kappa_3 = \mu = \kappa_4 = \kappa_5 = \dots .$$

In particular,

$$\kappa_6 - 2\kappa_5 + \kappa_4 = \mu - 2\mu + \mu = 0.$$

An estimate of $\beta \equiv \kappa_6 - 2\kappa_5 + \kappa_4$ is $b = k_6 - 2k_5 + k_4$, where these k-statistics are given in Kendall and Stuart (1969), page 280, (12.28). Let $s^2(b)$ be a consistent estimator for $\sigma^2(b)$. At level α one rejects the normal-plus-Poisson hypothesis if $|b/s(b)| > z(\alpha/2)$. The quantities $\mathrm{Var}(k_4)$, $\mathrm{Var}(k_5)$, $\mathrm{Var}(k_6)$, $\mathrm{Cov}(k_4,k_5)$, and $\mathrm{Cov}(k_4,k_6)$ needed to compute $\mathrm{Var}(b)$ are given in Kendall and Stuart(1969), pages 290-294. [Unfortunately, $\mathrm{Cov}(k_5,k_6)$ is not given.] These formulas are complicated but could be simplified, using (5.3), to provide an expression for $\sigma^2(b)$ under the normal-plus-Poisson hypothesis. This expression will involve only $E(Y)$, $\mathrm{Var}(Y)$, and $E(Z)$, which could be estimated unbiasedly by formulas provided by Sclove and Van Ryzin (1971). These estimates could then be substituted into the expression for $\sigma(b)$ to provide the required consistent estimate $s(b)$.

An alternative approach is subsampling. One partitions the sample into several (say, t) disjoint subsamples and computes an estimate b from each. Let b_j, $j = 1,2,\dots,t$, denote the subsample values. Define

$$\bar{b} = \Sigma_{j=1}^{t} b_j/t,$$

and take

$$s^2(\bar{b}) = \Sigma_{j=1}^{t}(b_j - \bar{b})^2/[t(t-1)] .$$

Then the test statistic is $\bar{b}/s(\bar{b})$. One needs to take t large enough so that approximate normality of $\bar{b}$ can be used.

6. Testing for normality, given normal-plus-Poisson. The presence of the Poisson component Z could affect adversely the power of the test of normality of Section 4. Accordingly, it makes sense to consider testing nested hypotheses in sequence. One first tests the hypothesis of Section 5, viz., $\kappa_6 - 2\kappa_5 + \kappa_4 = 0$. If this hypothesis is rejected, one stops and retains the full model. (The nature of the component Z is not then further specified.) On the other hand, if this hypothesis is accepted, one then tests the hypothesis

$$\kappa_4 = 0, \text{ given that } \kappa_6 - 2\kappa_5 + \kappa_4 = 0.$$

This is logically equivalent to the hypothesis

$$\kappa_6 - 2\kappa_5 = 0.$$

The statistic $c = k_6 - 2k_5$ is an unbiased estimator for $\kappa_6 - 2\kappa_5$. A test statistic is $c/s(c)$, where $s(c)$ is a consistent estimator for $\sigma(c)$. Either of the approaches of Section 5 could be used. The hypothesis of normality would greatly simplify the expression for $\sigma(c)$.

7. Testing for a Poisson distribution. Though we have focused on continuous r.v.'s X, it is of interest to note how one can test the hypothesis that X is Poisson (i.e., Y is zero and Z is *ordinary* Poisson). The r.v. X is Poisson if and only if $K(x)$ can have a jumpt only at $x = 1$. This corresponds to

$$\begin{aligned} 0 &= \int (x-1)^2 dK(x) \\ &= \int x^2 dK(x) - 2\int x dK(x) + \int dK(x) \\ &= (d^4/du^4) \log f(u)|_{u=0} - 2(d^3/du^3) \log f(u)|_{u=0} + (d^2/du^2) \log f(u)|_{u=0} \\ &= \kappa_4 - 2\kappa_3 + \kappa_2 \\ &= \partial , \end{aligned}$$

say. Let $d = k_4 - 2k_3 + k_2$ be the k-statistic estimate of ∂. Then a test statistic for the hypothesis of a Poisson distribution is $d/s(d)$, where $s(d)$ is a consistent estimator for $\sigma(d)$. Again, an alternative approach is provided by subsampling.

REFERENCES

[1] Borges, R. (1966). A characterization of the normal distribution (a note on a paper by Kozin). *Z. Wahrscheinlichkeitstheorie und Verew. Geite.* 5 244-246.

[2] DuMouchel, W.H.(1973). Stable distributions in statistical inference: 1. Symmetric stable distributions compared to other long-tailed distributions. *J. Amer. Statist. Assoc.* 68 469-477.

[3] DuMouchel, W.H.(1975). Stable distributions in statistical inference: 2. Information from stably distributed samples. *J. Amer. Statist. Assoc.* 70 386-393.

[4] Fama, E.F. and Roll, R.(1971). Parameter estimates for symmetric stable distributions. *J Amer. Statist. Assoc.* 66 331-338.

[5] Kendall, M.G. and Stuart, A.(1969). *The Advanced Theory of Statistics, Vol. I* (3rd edition). Hafner, New York.

[6] Loève, M.(1963). *Probability Theory* (3rd edition). Van Nostrand, New York.

[7] Pierre, P.(1967). Properties of non-Gaussian, continuous parameter, random processes as used in detection theory. Doctoral dissertation, The Johns Hopkins University.

[8] Pierre, P.(1969). New conditions for central limit theorems. *Ann. Math. Statist.* 40 319-321.

[9] Pierre, P.(1971). Infinitely divisible distributions, conditions for independence, and central limit theorems. *J. Math. Analysis and Applications* 33 341-354.

]10] Sclove, S.L. and Van Ryzin, J.(1971). Estimating the parameters of a convolution. *J. Royal Statist. Soc.(B)* 31 181-191.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
BOX 4348
CHICAGO, ILLINOIS 60680

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (18) AFOSR (19) TR-76-1258

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): (6) INFINITELY DIVISIBLE DISTRIBUTIONS IN STATISTICAL INFERENCE: HEAVY-TAILED DISTRIBUTIONS AND CONVOLUTION MODELS,

5. TYPE OF REPORT & PERIOD COVERED: ~~Interim~~

6. PERFORMING ORG. REPORT NUMBER: (15) AF-AFOSR-3050-76

7. AUTHOR(s): (10) Stanley L. Sclove

8. CONTRACT OR GRANT NUMBER(s): AFOSR 76-3050

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
University of Illinois at Chicago Circle
Department of Mathematics
Box 4348, Chicago, Illinois 60680

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F (16) 9769 (17) 05

11. CONTROLLING OFFICE NAME AND ADDRESS:
Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE: November 1, 1976

13. NUMBER OF PAGES: 9 + ii pp. (12) 13p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (11) 1 Nov 76

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Infinite divisibility, normality, Poisson distribution, cumulants, convolutions

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The family of infinitely divisible distributions is shown to provide alternative formulations in several inferential situations. In particular, the family provides heavy-tailed distributions and distributions for use in models involving convolutions, such as signal-plus-noise models. Characterizations of subfamilies of the infinitely divisible family are used to obtain statistical tests of membership in those sub-families. Special attention is given to the normal and normal-plus-Poisson sub-families.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 409950

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)